UNIT 3

PREPARATION METHODS FOR SEQUENCING (FROM PPTs)

ADAPTORS

Adapters are short DNA sequences that are ligated onto the ends of DNA inserts (target sequences) during library preparation for next-generation sequencing (NGS). They play a crucial role in ensuring the compatibility of the sequencing library with the sequencing platform and facilitating the sequencing process.

The **basic structure** of an NGS adapter includes several key functional motifs-

1. **Flow Cell Binding Sequences (P5/P7) –** Bind library to sequencing platform.
2. **Index Sequences (i5/i7) –** Barcodes to identify samples during multiplexing.
3. **Sequencing Primer Binding Sites (Rd1/Rd2 SP) –** Initiate sequencing process.
4. **Unique Molecular Identifiers (UMI) –** Additional barcodes for individual target sequences within a sample.

They can also include additional sequence elements that confer additional functions, such as sample indexes and UMIs, to improve library analysis

Types of Adapters –

1. **Based on their index positions –** Single-indexed, Dual-indexed → Dual-indexed provides better multiplexing.
2. **Based on their compatibility with PCR – free library preparation – Full, Partial adapters**
3. **Unique Dual Indexes (UDIs) –** Reduce index-hopping errors.

Applications –

Used in all NGS library preparation, including short- and long-read sequencing, DNA and RNA sequencing.

INDEX

Indexing allows for the identification of individual samples within the sequencing data by attaching unique sequences to each sample, enabling the separation of reads from different samples during analysis.

1. **Essential for Multiplexing –** Allows combining multiple samples in one run, saving time and money.
2. **Sample Identification –** Unique indexes attached to each sample enable separation of reads during analysis.
3. **Error Correction –** Helps identify and remove misassigned reads caused by errors like index hopping.

Types of Indexing –

1. **Inline Indexing –** Index sequences directly added to sequencing adapters.
2. **Multiplex Indexing –** Index sequences introduced during adapter ligation or PCR amplification.

Challenges –

1. **Index Hopping –** Data from one sample contaminates another's index. (Mitigated by dual indexes & advanced library prep kits).
2. **Index Cross-Talk –** Overlapping index sequences lead to misassigned reads. (Reduced by unique molecular identifiers (UMIs) & bioinformatics tools).

Bioinformatics Solutions –

1. **Deconvolution Algorithms –** Remove adapter dimers and identify index hopping instances.
2. **Unique Molecular Identifiers (UMIs) –** Further reduce index-hopping errors by tagging individual target sequences.

LIBRARY PREPARATION METHODS

1. Bridge amplification

Bridge amplification is a process used in Illumina sequencing that involves preparing genomic DNA samples by fragmenting and ligating them to adapters before attaching them to a flow cell surface. The process involves several steps:

1. **Fragmentation:** The genomic DNA is fragmented into smaller pieces, typically using enzymes like DNase or sonication.
2. **Adapter Ligation:** The fragmented DNA is then ligated to adapters, which are short DNA sequences that facilitate the sequencing process. The adapters contain specific sequences that allow them to bind to the flow cell surface.
3. **Bridge Amplification:** The adapter-ligated DNA is then amplified on the surface of the flow cell using bridge amplification. This process involves the following steps:
    (a) **Primer Binding:** Primers on the surface of the flow cell bind to the adapter sequences.
    (b) **Unlabeled Nucleotides and Polymerase:** Unlabeled nucleotides and polymerase enzymes are added to the surface, allowing the synthesis of new strands that become double-stranded.
    (c) **Washing and Repeating:** The original strands are washed away, and the process is repeated to amplify multiple copies of each fragment in parallel.

Bridge amplification is essential for Illumina sequencing as it allows for the generation of millions of copies of each DNA fragment, which are then sequenced on the flow cell surface. This process ensures that each fragment is amplified multiple times, resulting in high-quality sequencing data.

2. Emulsion PCR

Overview of Emulsion PCR

- Emulsion PCR is a technique used in NGS to amplify DNA sequences attached to beads.
- It involves compartmentalizing DNA fragments with primer-coated beads into water-in-oil emulsion droplets, where each droplet contains one DNA fragment.
- The droplets then act as individual PCR reactors to amplify each fragment clonally onto a single bead.
- After thermal cycling, millions of copies of the DNA fragment are attached to each bead for downstream sequencing applications.

Applications of Emulsion PCR in NGS

1. **Library Enrichment:** Emulsion PCR is widely used for library enrichment in NGS, providing high-quality and high-yield NGS libraries for sequencing.
2. **Single-Cell Genomics:** Emulsion PCR can effectively amplify individual fragments of a single cell's entire genomic DNA, enabling studies on cell-to-cell heterogeneity.
3. **Rare Variant Detection:** Emulsion PCR is used for the selective detection of rare variants by partitioning low-abundance template DNA into individual emulsion droplets for amplification.
4. **Forensic Analysis:** Emulsion PCR is employed in forensic analysis for the detection of trace amounts of DNA obtained from crime scenes.

Advantages of Emulsion PCR

1. Reduces PCR duplications by amplifying each DNA fragment into individual reactions within emulsion droplets.
2. Provides high-throughput amplification of thousands to millions of DNA fragments in a single reaction.
3. Highly sensitive and accurate in amplifying even a few template DNA molecules present in the sample.
4. Cost-effective, simple, and fast method for amplifying DNA fragments for various sequencing applications.

Challenges and Limitations

1. Requires precise optimization of PCR conditions to ensure efficient and uniform amplification.
2. Emulsion stability can be affected by temperature fluctuations during the PCR process.

In summary, emulsion PCR is a crucial library preparation technique used in NGS to generate high-quality, high-yield sequencing libraries by compartmentalizing and amplifying individual DNA fragments within emulsion droplets.

DATA FORMATS

Next-generation sequencing (NGS) generates vast amounts of data, which are stored in various file formats. These formats are used to store different types of data, including nucleotide sequences, quality scores, alignments, genomic annotations, and coverage data.

1. FASTQ Format

The FASTQ format is a text-based format used to store both biological sequences (usually nucleotide sequences) and their corresponding quality scores. Each sequence in a FASTQ file consists of four lines:

1. **Sequence Identifier:** Starts with an '@' symbol and includes information about the sequencer, flowcell, and position on the flowcell.
2. **Sequence:** The raw sequence data.
3. **Separator Line:** Starts with a '+' symbol and can include the same sequence identifier.
4. **Quality Scores:** Encoded as ASCII characters, each representing the likelihood of a sequencing error at that particular base in PHRED-scores.

Importance

1. **Wide Adoption:** FASTQ is the most widely used format in sequence analysis.
2. **Standardization:** The format is standardized, making it easy to share and analyze data across different platforms.
3. **Quality Control:** FASTQ files include quality scores, which are essential for quality control and error detection in sequencing data.
4. **Generation:** FASTQ files are generated by sequencing machines and can be converted from other formats like BAM and SFF.
5. **Analysis:** FASTQ files are used as input for various downstream analysis tools, including aligners, assemblers, and quality control tools.

2. Subreads Format

Subreads is a data format used in Next-Generation Sequencing (NGS), specifically for storing data from Oxford Nanopore Technologies (ONT) sequencing platforms.

- **Purpose:** Subreads are used to store the raw signal data and the basecalled sequence information generated by ONT sequencing instruments.
- **Format:** Subreads are stored in a binary format, which is optimized for efficient storage and processing of the large amounts of data produced by ONT sequencing.
- **Composition:** Each subread contains the following information:
    1. The raw electrical signal data measured by the nanopore as the DNA molecule passes through.
    2. The basecalled sequence, which is the nucleotide sequence determined from the raw signal data.
    3. Additional metadata, such as the start and end positions of the subread within the original DNA molecule.

- **Relationship to Reads:**
  ONT sequencing: Single DNA molecule gets sequenced multiple times (subreads). Each of these repeated sequencing events is called a "subread".
  Subreads are used to create high-quality ‘reads of insert’ (ROI) by combining information from multiple passes.

- **Applications:**
    1. **Genome assembly:** Long subreads improve assembly of complex genomes.
    2. **Structural variation detection:** Helps identify large-scale variations.
    3. **Epigenetic modifications:** Raw data helps detect methylation and other modifications.

- **Tools and Software:** Specialized tools, such as those provided by ONT, are used to work with subread data, including basecalling, alignment, and downstream analysis.

3. Nanopore data

Nanopore sequencing data is generated by Oxford Nanopore Technologies (ONT) sequencing platforms and has some unique characteristics compared to other Next-Generation Sequencing (NGS) data formats:

1. **Raw Signal Data (FAST5):**
    - Stores electrical signals from DNA/RNA passing through the nanopore.
    - Binary format, allows extracting specific information like raw data and basecalled sequence.

2. **Basecalled Sequence Data (FASTQ):**
    - Processed from FAST5 data to get the underlying nucleotide sequence.
    - Standard FASTQ format with sequence and quality scores.
    - May have higher error rates and homopolymer errors compared to other NGS platforms.

3. **Subreads:**
    - Repeated sequencing events of a single DNA/RNA molecule.
    - Stored in a proprietary binary format.
    - Contain raw signal data and basecalled sequence for each event.
    - Consensus sequence of subreads is called the "read of insert" (ROI).

4. **Alignment and Analysis Formats:**
    - Standard formats like SAM/BAM/CRAM for alignment to reference genomes (minimap2, graphmap).
    - Additional formats like BED/VCF for genomic features, variants, or analysis results.

Key Point: Nanopore data formats handle unique characteristics like long reads, high error rates, and raw signal storage. Understanding these formats is crucial for effective analysis.

4. Single cell data

Single-cell data in Next-Generation Sequencing (NGS) refers to the analysis of individual cells, which involves the sequencing of the entire genome or transcriptome of a single cell. This approach allows researchers to study cellular heterogeneity and identify specific cell types, subpopulations, and their roles in complex biological systems.

Single-cell data is typically stored in various file formats, including:

1. **FASTQ:** Stores both nucleotide sequences and their corresponding quality scores.
2. **BAM:** Binary Alignment/Map format, which contains the alignment information and additional metadata.
3. **CRAM:** Compression Alignment/Map format, which further compresses the alignment information by storing only the differences between the aligned sequences and a reference sequence.
4. **BED/GTF:** Browser Extensible Data and Gene Transfer Format, used for storing gene and feature annotations.
5. **bedgraph:** Used to store continuous-valued data, such as gene expression levels or coverage data across the genome.

SINGLE-END VS PAIRED-END VS MATE-PAIR READS

Criteria	Single-End Reads	Paired-End Reads	Mate-Pair Reads
Read Structure	Single sequence read from one end of a DNA fragment	Two sequence reads from opposite ends of a DNA fragment	Two sequence reads from separate, distant locations on the same DNA molecule
Fragment Size	Variable, typically short (35-500 bp)	Variable, typically short (50-400 bp)	Larger and more variable (2-10 kb)
Insert Size	N/A	Distance between the sequenced ends of the fragment (adjustable during library preparation)	Distance between the two sequenced regions (defined by library preparation method)
Paired Information	No information about the other end of the fragment	Information about the other end (sequence and distance)	Information about the relative location on the same molecule, but not necessarily the sequence
Library Preparation	Simpler, fragments are randomly sheared and sequenced from one end	More complex, fragments are sequenced from both ends	More complex, involves circularization of DNA and fragmentation
Applications	Suitable for basic tasks like SNP detection, gene expression analysis	Useful for de novo assembly, indel detection, and variant calling	Effective for resolving complex rearrangements, large insertions, and scaffolding genomes
Cost	Relatively inexpensive	More expensive than single-end	More expensive than paired-end
Throughput	Higher due to simpler library preparation	Lower than single-end due to sequencing both ends	Lower than paired-end due to larger fragment size
Coverage	Lower coverage for repetitive regions	Higher coverage for repetitive regions and complex regions	May require additional library preparation steps for even coverage
Error Correction	May require more error correction due to shorter read lengths	Benefits from paired-end information for error correction	May require additional strategies for error correction due to larger fragment size

NGS DATA SOURCES

Next-Generation Sequencing (NGS) data sources are repositories that store and manage large amounts of sequencing data generated from various platforms. These data sources are crucial for bioinformatics analysis and research. Here is an overview of the key NGS data sources:

1. **NCBI (National Center for Biotechnology Information)**
   - **Sequence Read Archive (SRA):** A publicly available repository of high-throughput sequencing data, part of the International Nucleotide Sequence Database Collaboration (INSDC).
   - **Data Processing, Status, and Release:** SRA accepts data from all kinds of sequencing projects, including clinically important studies. Data are subject to automated and manual processing to ensure data integrity and quality before being made available to the public.

2. **EBI-ENA (European Bioinformatics Institute-European Nucleotide Archive)**
   - **European Nucleotide Archive (ENA):** A repository for the world public domain nucleotide sequence data output.
   - **Data Content and Scope:** ENA covers a spectrum of data types, including raw reads, assembly data, and functional annotation.
   - **Assembly Information Services:** ENA provides services for genome assembly information and has faced significant growth in genome assembly submission rates and data volumes.

3. **DDBJ-SRA (DNA Data Bank of Japan-Sequence Read Archive)**
   - **Sequence Read Archive (SRA):** A repository of high-throughput sequencing data, part of the International Nucleotide Sequence Database Collaboration (INSDC).
   - **Data Exchange:** DDBJ-SRA exchanges data with other INSDC members, including NCBI and EBI-ENA.

4. **GEO (Gene Expression Omnibus)**
   - **Gene Expression Omnibus (GEO):** A public functional genomics data repository supporting the submission and sharing of microarray, RNA-seq, and other high-throughput sequencing data.
   - **Data Types:** GEO accepts various data types, including microarray, RNA-seq, and other high-throughput sequencing data.

Importance of NGS Data Sources

1. **Data Sharing:** NGS data sources facilitate data sharing and collaboration among researchers, enabling the advancement of scientific knowledge.
2. **Data Analysis:** These repositories provide a platform for bioinformatics analysis and research, allowing for the development of new tools and methods.
3. **Data Quality Control:** NGS data sources ensure data integrity and quality through automated and manual processing, ensuring the reliability of the data.

SEQUENCE QUALITY MEASURES, PHRED QUALITY SCORE, FASTQC, TRIMMOMATIC, FASTX – TO BE STUDIED FROM PPT

REST ALL TOPICS – STUDY FROM PPT